9-1-2019

Virtual Audits of Streetscapes by Crowdworkers

Tomoya Hanibuchi
Portland State University, hanib2@pdx.edu

Tomoki Nakaya
Tohoku University

Shigeru Inoue
Tokyo Medical University

Citation Details

Hanibuchi, T., Nakaya, T., & Inoue, S. (2019). Virtual audits of streetscapes by crowdworkers. Health & place, 59, 102203.

Health and Place 59 (2019) 102203

Contents lists available at ScienceDirect

Health & Place

journal homepage: www.elsevier.com/locate/healthplace

Virtual audits of streetscapes by crowdworkers

Tomoya Hanibuchi[a,b,*], Tomoki Nakaya[c], Shigeru Inoue[d]

[a] School of International Liberal Studies, Chukyo University, 101-2 Yagoto-honmachi, Showa-ku, Nagoya, 466-8666, Japan
[b] Department of Geography, Portland State University, Portland, OR, USA
[c] Graduate School of Environmental Studies, Tohoku University, Sendai, Miyagi, Japan
[d] Department of Preventive Medicine and Public Health, Tokyo Medical University, Tokyo, Japan

ARTICLE INFO

Keywords:
Audits
Neighborhood walkability
Google street view
Crowdsourcing

ABSTRACT

Audits have been used to provide objective ratings of neighborhood environments. Physical audits, however, are time- and resource-intensive. This study examines the efficiency and reliability of virtual auditing using Google Street View and crowdsourcing to conduct walkability audits of streets in Japan. Overall, 830 street segments were physically and virtually audited by two trained auditors; 300 untrained crowdworkers also virtually audited 3 street segments. Statistical analysis found good inter-source and inter-rater reliability. This study helps establish crowdsourced virtual auditing as a valuable method of measuring neighborhood walkability, reducing audit costs as well as enabling large-scale auditor recruitment while maintaining reliability.

1. Introduction

Over the past several decades, researchers have studied the relationship between neighborhood built environments and health behavior, for instance, walking and physical activity (Ding and Gebel, 2012; Sallis et al., 2016). Neighborhood walkability, or pedestrian-friendly built environments that support walking, are attributed to be positively associated with its amount and frequency (Ferdinand et al., 2012; Grasser et al., 2013; Saelens and Handy, 2008). Although most studies measured neighborhood walkability using residents' perceptions or Geographic Information Systems (GIS) measures, these approaches suffer from methodological challenges such as same-source bias (Diez Roux, 2007) and limited data on neighborhood conditions.

Many researchers have turned to audits, or systematic social observation, to fill this methodological gap (Brownson et al., 2009; Schaefer-McDaniel et al., 2010). Audits provide objective ratings of neighborhood environments, including micro-scale elements, such as conditions of sidewalks. One of the biggest challenges of this method is that it is highly time- and resource-intensive; therefore, the study area needs to be substantially narrow and close to the auditors' homes (Kelly et al., 2013; Rundle et al., 2011). To address this issue, a growing number of studies conducted virtual audits using street-level imagery from Google Street View (GSV) instead of in-person/physical audits because they cost less, are easier to conduct, and save time (Rzotkiewicz et al., 2018). Studies have repeatedly established virtual audits as an efficient and reliable method, providing strong inter-source (Badland et al., 2010; Ben-Joseph et al., 2013; Clarke et al., 2010; Malecki et al., 2014; Pliakas et al., 2017; Rundle et al., 2011) and inter-rater reliability (Kelly et al., 2013; Pliakas et al., 2017).

However, virtual audits are not sufficient to reduce costs because it still relies heavily on manual work, and auditors are generally requested to spend more time training in order to accurately rate features of streetscapes. This may result in a limited number of trained personnel to conduct audits, thus restricting the area of study. Further progress in reducing time costs and improving the recruitment of higher numbers of auditors is necessary to expand the audit area. One possibility to address this issue is the use of a computer vision approach. Automated characterizations of the neighborhood built environments can lower study costs and enable researchers to cover larger geographic areas (Nguyen et al., 2018). However, the range of items audited using automated methods is still very limited. Therefore, to fill the gap between a trained auditor approach and computer vision approach, an intermediate method must be developed (Hipp et al., 2017).

We propose the use of paid crowdsourcing as a complement to virtual auditing. By combining virtual audits with crowdsourcing, many auditors can participate in the observation process, with the audit area expanded to cover more number of streets. Nevertheless, research applying this approach to neighborhood studies is limited (Hara et al., 2013; Hipp et al., 2017). In one study, Hara et al. (2013) investigated the feasibility of using untrained crowdworkers (n = 402) recruited from Amazon Mechanical Turk to find, label, and assess sidewalk accessibility problems in GSV imagery, showing that crowdworkers were

* Corresponding author. School of International Liberal Studies, Chukyo University, 101-2 Yagoto-honmachi, Showa-ku, Nagoya, 466-8666, Japan.
E-mail address: info@hanibuchi.com (T. Hanibuchi).

https://doi.org/10.1016/j.healthplace.2019.102203
Received 10 May 2019; Received in revised form 19 August 2019; Accepted 5 September 2019
Available online 11 September 2019

capable of determining the presence of an accessibility problem with 81% accuracy. This study suggests that the above method can also be applied for walkability audits.

This study examines the efficiency and reliability of audits conducted via crowdsourcing. In order to conduct a virtual audit using untrained crowdworkers, a simple and easy-to-use checklist as well as brief instructions are required. These materials may not require long hours of intensive training to review, which is difficult to implement when recruiting large numbers of crowdworkers. Several researchers have developed simple audit tools with approximately 15 items (Hoedl et al., 2010; Sallis et al., 2015). For example, MAPS-Mini was designed to be short enough to use in practice, accompanied by limited training for observers (Sallis et al., 2015). With reference to previous studies, we followed a three-step process. First, we prepared a simple checklist and a brief context-specific instruction. Next, we tested the efficiency and reliability of the checklist and instructions using the data of two trained auditors' physical and virtual audits. Finally, 300 untrained crowdworkers' audit ratings were examined after conducting a virtual audit of three sample street segments in Nagoya, Japan, in terms of inter-rater reliability and differences in the ratings as a function of individual attributes and experience using GSV.

This research contributes to existing neighborhood studies in two significant ways. First, we explore the utility of audits when conducted by a large number of untrained crowdworkers, examining the nature and extent to which their ratings differ from trained auditors, and whether there are any systematic factors shaping the ratings of crowdworkers, such as demographic characteristics, place of residence, and experience of using GSV. In addition, we apply the virtual audit methodology to an East Asian city, expanding the body of existing research on audits to test its applicability to non-Western countries as well. Studies using virtual audits are still limited to the United States and other Western countries (Rzotkiewicz et al., 2018), with studies in other geographical contexts being required to explore the method's generalizability. Japanese cities are considered suitable cases for such empirical exploration, as they have different urban forms than that of Western countries (Kaido, 2006; Shelton, 2012), and the coverage of GSV is very high.

2. Methods

2.1. Audit instructions and walkability checklist

The instructions and checklist to assist crowdworkers were intentionally crafted to be simple and easy to understand as they were not subject to training from experienced auditors. Such simple and easy to understand tools have not been tested in Japan. We created brief, one-page visual instructions for crowdworkers (Fig. 1). To develop the audit checklist, we first reviewed some of the existing audit tools such as WASABE (Malecki et al., 2014), PEDS (Clifton et al., 2007), and MAPS-mini (Sallis et al., 2015), with a specialized focus on those developed in Asian contexts (CUBEST: Su et al., 2014, EAST-HK: Cerin et al., 2011). We then selected a limited number of items that were (1) often used in previous audit tools (Cerin et al., 2011; Clifton et al., 2007; Sallis et al., 2015), (2) covering various aspects of micro-scale streetscape, and (3) considered suitable to the study setting (Nagoya city, Japan) while reducing the number of items. Items such as presence of wide sidewalk, heavy traffic, crosswalk, streetlights, street trees, attractive streetscape, graffiti and litter, and abandoned buildings were included as basic elements for micro-scale walkability, which covers aspects of physical conditions, safety, and aesthetics. These aspects basically corresponded to themes of EAST-HK (Cerin et al., 2011), but items that could be measured by GIS (e.g., "destinations" such as shops, parks, and stations) were not included because our checklist was designed to capture a micro-scale streetscape. Characteristics of the East Asian ultra-dense cities such as crowdedness, the presence of man-made obstacles to walking such as cars parked on footpaths, and steep terrain (Cerin et al., 2011) were also included in the checklist. Another addition to the checklist was the presence of a traffic mirror which is often installed at blind corners, intersections, and entrances to improve visibility of pedestrians and cars for safety purposes, considering the low visibility due to narrow streets with crowded buildings being a characteristic of streetscape in Japanese cities (Fig. 2).

After testing the initial checklist on randomly selected sample streets, we repeatedly revised the checklist and instructions in order to increase the coverage of the various aspects of neighborhood environments while maintaining the audit tools' simplicity and ease of use. Ultimately, we developed a simple checklist consisting of 14 items. In order to make the checklist easy to administer by even untrained auditors, all 14 items were dichotomous (i.e., present or absent).

2.2. Audits by trained auditors

Between May and June 2018, two trained auditors (A and B) independently conducted virtual audits using the checklist. The auditors were freelance research assistants, whose demographics were similar in terms of age and gender. They conducted audits online: start by clicking the designated URLs that linked to the starting points and then "walk-through" the target streets on GSV. Most of the GSV images were captured in 2017 (40.1%) and 2016 (53.9%), while a few were captured before or in 2015 (6.0%). One of the auditors (A) also conducted an in-person audit of the same street segments using the same checklist between April and May 2018. Initial instructions, pre-testing, and consultation by one of the authors were performed prior to the audit. Overall, 20 neighborhoods (*chocho-aza*) were randomly selected in Nagoya City, the center of the third largest metropolitan area of Japan, and all street segments within the neighborhoods were audited. The auditors were instructed to audit the streets' right and left sides separately. Both sides of 415 streets were audited, with all 830 street segments totaling 74.8 km in length.

2.3. Audits by crowdworkers

We recruited 300 crowdworkers using *Lancers*, one of the largest paid crowdsourcing platforms in Japan. The demographics of crowdworkers registered with *Lancers* indicate diverse regional representation (Kavanagh et al., 2016). On September 6, 2018, we posted a "task" on *Lancers* seeking auditors to conduct a GSV audit of three street segments. Estimated time for completing the task, which included reading the instructions and answering a short questionnaire, was about 30 min, and compensation for the task was set at 500 Japanese Yen (approximately 5 USD).

After accepting the task, crowdworkers were invited to review the visual instructions before conducting a GSV audit for the target street segments using the 14-item checklist. The three street segments were selected from the 830 street segments audited in Nagoya City, with each chosen based on the highest, middle, and the lowest walkability ratings observed by the trained auditors to maximize variation in street segments. These segments were rated exactly the same by the two trained auditors. After completing the audit, crowdworkers were also requested to answer a short questionnaire regarding their demographics, residential settings, work environment, and experience of using GSV.

2.4. Analysis of efficiency and reliability

Using data from trained auditors, the efficiency of virtual audits compared to in-person audits was assessed by calculating the average time taken for auditing one street segment. Then, we calculated the prevalence of each item and scores by aggregating the items. Prevalence of each item was calculated as a proportion of street segments that were rated as being present (e.g., presence of sidewalks). Sub-total scores were calculated by summing the number of present items according to different aspects of micro-scale streetscape: physical

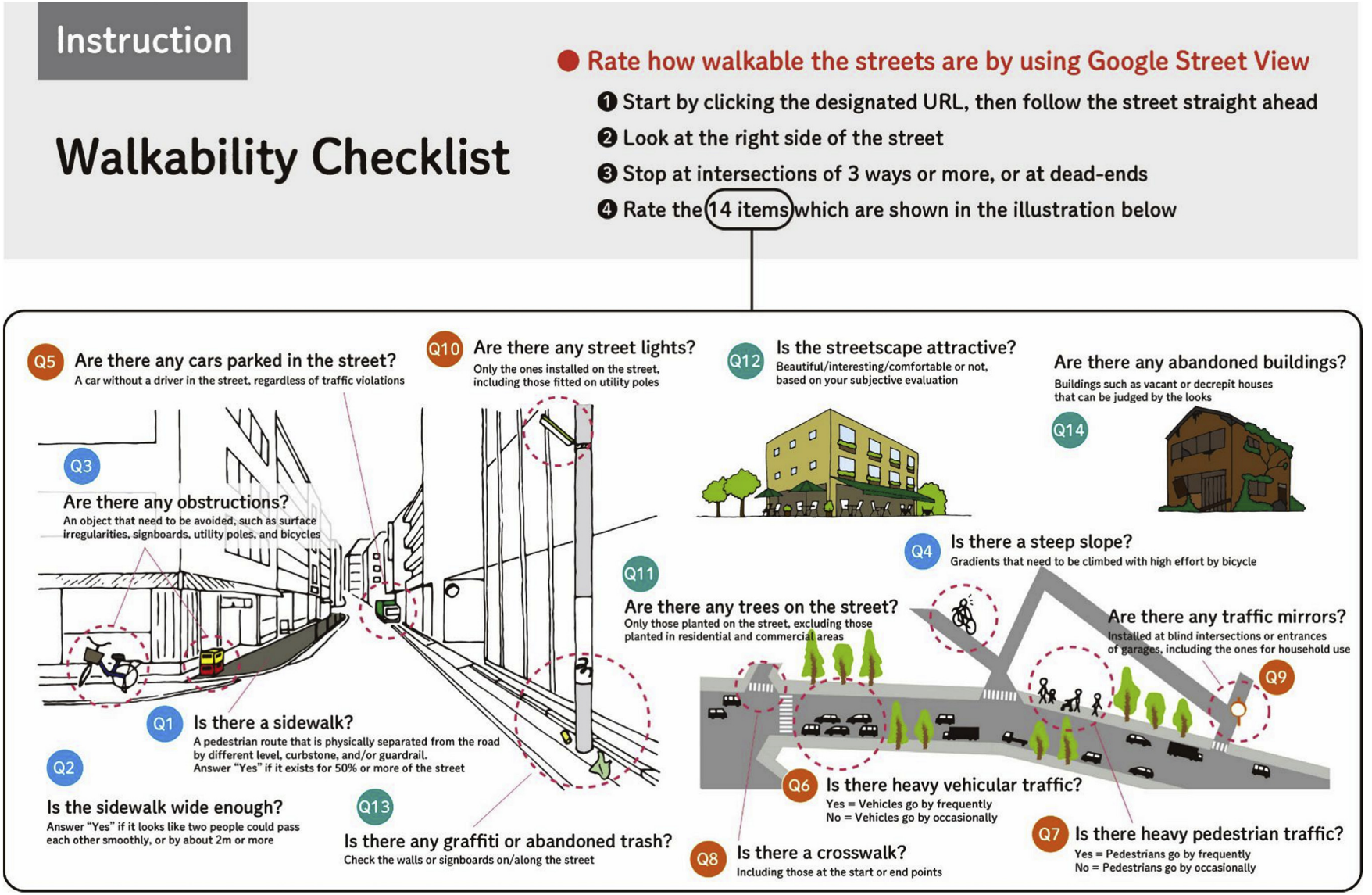

Fig. 1. Brief visual instructions explaining the walkability checklist.

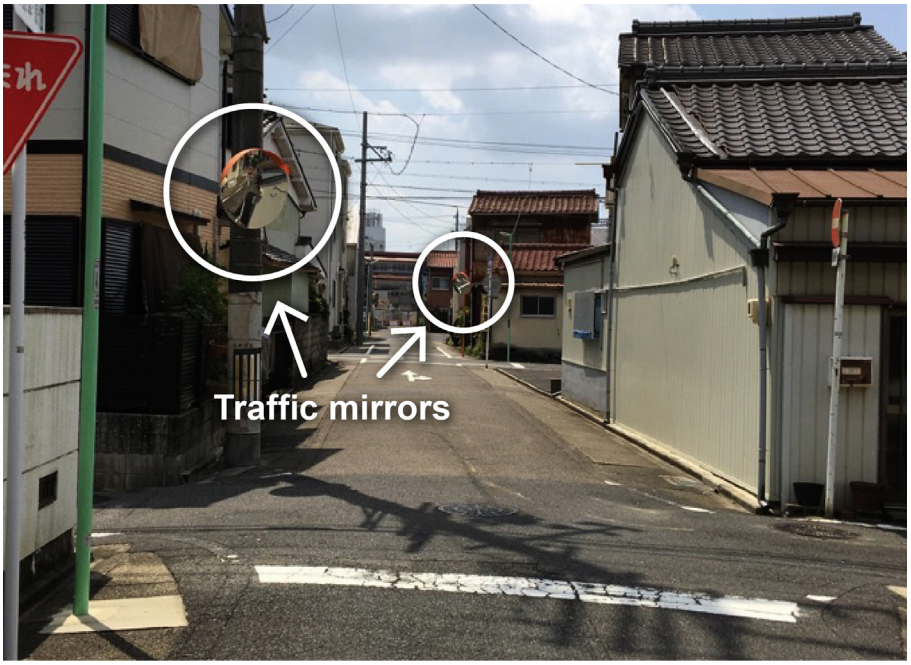

Fig. 2. Example of neighborhood streetscape of the study area including traffic mirrors.
Source: Author's photo.

conditions score consists of sidewalks, wide sidewalks, obstructions, and steep slopes (Q1, 2, 3, and 4 in Fig. 1); safety score consists of street parking, heavy traffic, heavy foot traffic, crosswalks, traffic mirrors, and streetlights (Q5, 6, 7, 8, 9, and 10 in Fig. 1); and aesthetics score consists of street trees, attractive streetscape, graffiti and litter, and abandoned buildings (Q11, 12, 13, and 14 in Fig. 1). Items that were assumed to be non walking-friendly environment (i.e., obstructions, steep slopes, street parking, heavy traffic, heavy foot traffic, traffic mirrors, graffiti and litter, and abandoned buildings) were reverse coded before summation. Therefore, it can be interpreted that the higher the scores, the more walkable the street. Total scores were calculated by summing all the sub-total scores.

To assess reliability of the ratings through the strength of agreement, percentage of agreement and Kappa coefficients for each item were calculated between in-person and virtual audits (i.e., inter-source reliability) and between the two trained auditors (i.e., inter-rater reliability). Percentage of agreement values were interpreted in the following manner: above 90% (excellent), 80–89% (very good), 70–79% (good), 60–70% (moderate) less than 59% (poor) (Malecki et al., 2014). Kappa coefficient values between 0.81 and 1 were interpreted as almost perfect, between 0.61 and 0.80 as substantial, 0.41 and 0.6 as moderate, and below 0.40 as poor-to-fair (Landis and Koch, 1977). Intraclass correlation coefficients (ICC) for total and sub-total scores were used to further assess inter-rater reliability. ICC values less than 0.39, between 0.40 and 0.59, between 0.60 and 0.74, and between 0.75 and 1.00 were considered indicative of poor, fair, good, and excellent reliability, respectively (Cicchetti, 1994).

For the crowdworkers' audit, the number of days required for recruiting crowdworkers and average time for completing the task were reported to assess the efficiency of this method. In order to examine the reliability of untrained crowdworkers' ratings, the percentage of agreement between the ratings of crowdworkers and the trained auditors was calculated. Additionally, ICCs were calculated for the sub-total and total scores between ratings by crowdworkers and the trained auditors. We also analyzed the percentage of agreement as a function of crowdworkers' age, gender, residential settings, work environments, and experience using GSV. If no or only small differences were found, the results may indicate that many crowdworkers can potentially participate in a virtual audit regardless of their attributes.

Statistical analyses were performed using IBM SPSS Statistics 24.

3. Results

3.1. Virtual and in-person audits by trained auditors

The average time for auditing one street segment was 2.72 min (± 1.45 min) for in-person audits and 1.83 min (± 0.98 min) and 1.36 min (± 0.56 min) for the virtual audits conducted by the two auditors, respectively. The virtual audits took between two-thirds to one-half of the time spent conducting in-person audits. The longer the street segments, the less time it took using virtual methods. In-person audits for street segments 300 m or longer took an average of 3.25 min longer than virtual audits conducted by the same auditor, with the time difference reduced to an average of 0.44 min for street segments shorter than 50 m. In-person audits required more time and incurred more travel expenses for the auditor to visit various streets located across the city. Importantly, virtual audits could be conducted regardless of weather conditions and the time of the day.

Table 1 presents a prevalence of the 14 items on the audit checklist, and Table 2 shows descriptive statistics of sub-total and total scores, by each trained auditor according to the audit method. Prevalence of the 14 items were generally similar between the methods and auditors, while some items showed large differences in the value, such as graffiti and litter between auditor A and auditor B. Sub-total and total scores were almost the same between the in-person and virtual audit by auditor A, although scores of auditor B were slightly higher for physical, aesthetic, and total scores compared to those of auditor A.

Table 3 shows the inter-source and inter-rater reliability scores. ICCs for total scores were interpreted as excellent in terms of inter-source reliability (0.84) and inter-rater reliability (0.75). Most subtotal scores also showed excellent to good reliability. ICCs for the physical condition score were also excellent: 0.89 for inter-source reliability and 0.78 for inter-rater reliability, while ICC measuring inter-rater reliability of aesthetic scores was only fair (0.40).

With regards to the Kappa coefficients of the 14 items, eight were categorized as having perfect or substantial inter-source reliability,

Table 1
Prevalence of 14 items rated by trained auditors.

No.	Items	In-person audit by auditor A (%)	Virtual audit by auditor A (%)	Virtual audit by auditor B (%)
1	Sidewalks	30.4	30.6	31.1
2	Wide sidewalks	20.5	13.7	25.2
3	Obstructions	78.1	84.8	59.2
4	Steep slopes	7.7	8.2	7.3
5	Street parking	7.1	16.1	16.4
6	Heavy traffic	17.7	7.8	11.9
7	Heavy foot traffic	5.2	2.8	7.6
8	Crosswalks	43.9	44.8	43.6
9	Traffic mirrors	19.6	15.2	12.3
10	Street lights	61.0	61.3	58.0
11	Street trees	14.2	14.1	13.9
12	Attractive streetscape	5.7	7.8	15.9
13	Graffiti and litter	21.7	22.9	1.7
14	Abandoned buildings	31.2	31.1	8.7

Table 2
Descriptive statistics of sub-total and total scores rated by trained auditors.

Scores	Definition[a]	In-person audit by auditor A		Virtual audit by auditor A		Virtual audit by auditor B	
		Mean	SD	Mean	SD	Mean	SD
Physical condition score	Σ[item 1–4]	1.65	1.12	1.51	0.94	1.90	1.25
Safety score	Σ[item 5–10]	4.55	0.80	4.64	0.85	4.53	0.86
Aesthetic score	Σ[item 11–14]	1.67	0.87	1.68	0.92	2.19	0.68
Total score	Σ[item 1–14]	7.87	1.82	7.83	1.87	8.63	1.91

[a] Items that were assumed to be non walking-friendly environments (i.e., obstructions, steep slopes, street parking, heavy traffic, heavy foot traffic, traffic mirrors, graffiti and litter, and abandoned buildings) were reverse coded before summation. It is assumed that the higher the scores, the more walkable the street.

Table 3
Inter-source and inter-rater reliability of walkability audit tools.

	Inter-source reliability (in-person and virtual)		Inter-rater reliability (between auditors)	
	Kappa/ ICC[a]	% agreement	Kappa/ ICC[b]	% agreement
Sidewalks	0.99	99.5	0.97	98.8
Wide sidewalks	0.76	93.3	0.64	88.6
Obstructions	0.59	87.5	0.38	73.0
Steep slopes	0.92	98.8	0.67	95.3
Street parking	0.30	85.4	0.83	95.4
Heavy traffic	0.49	88.4	0.64	93.5
Heavy foot traffic	0.40	95.4	0.44	94.5
Crosswalks	0.96	97.8	0.92	95.9
Traffic mirrors	0.76	93.1	0.68	92.3
Streetlights	0.85	92.7	0.77	88.9
Street trees	0.98	99.4	0.94	98.6
Attractive streetscape	0.58	94.7	0.27	84.5
Graffiti and litter	0.44	80.7	0.06	77.6
Abandoned buildings	0.68	86.1	0.29	75.7
Physical condition score	0.89		0.78	
Safety score	0.66		0.71	
Aesthetic score	0.73		0.40	
Total score	0.84		0.75	

[a] A two-way mixed model ICC, absolute agreement, single measures.
[b] A two-way random model ICC, absolute agreement, single measure.

with two items indicative of poor-to-fair reliability. Nine items had Kappa coefficient values indicative of perfect or substantial inter-rater reliability, with four items receiving values suggesting poor-to-fair reliability. For example, Kappa coefficients for sidewalks (0.99 and 0.97), crosswalks (0.96 and 0.92), and street trees (0.98 and 0.94) were shown to be very high for both inter-source and inter-rater reliability. Street parking (0.30) and heavy foot traffic (0.40), both of which tend to change over short periods of time (hourly or even by the minute), demonstrated poor-to-fair agreement between in-person and virtual audit scores. Items susceptible to subjective assessment also showed poor-to-fair agreement between the auditors; they included obstructions (0.38), attractive streetscape (0.27), graffiti and litter (0.06), and abandoned buildings (0.29). However, when looking at percentage of agreement, all of the reliability values were over 70%. Some items with low Kappa coefficients showed excellent-to-good percent of agreement scores due to the very low or high prevalence.

3.2. Crowdworker audits

Three hundred crowdworkers completed the audits within two days of posting the task. The average time for completing the task, including reviewing the visual instructions, auditing using GSV, and answering the short questionnaire was 19.4 min, about 10 min shorter than anticipated prior to conducting the study.

Tables 4 and 5 present the results of crowdworker audits compared to those conducted by the trained auditors. Table 4 shows the average percentage of agreement between the observations of the 300 untrained crowdworkers and the trained auditors, with the latter used as a gold standard. Overall, the average percentage of agreement for all items and all street segments was 83.7% (very good). The average value for all items did not vary significantly across the three street segments. However, percentages of agreement differed greatly across individual items. For example, the average value for heavy foot traffic of all three streets was 98.9% while only 51.6% for obstructions. Reliability of the sub-total and total scores between crowdworkers and trained auditors were calculated by ICC and is shown in Table 5. ICC for the total score was 0.65 (good), which was slightly lower when compared to that of between trained auditors (0.75). Similarly, sub-total scores were also shown to be lower by approximately 0.1 compared to those of between trained auditors.

Table 6 shows the differences of average percentage of agreement according to crowdworker demographics, residential statuses, working environments, and experience using GSV. Scores did not significantly vary by gender (range = 0.5), residential location (1.5), history of residence in Nagoya (1.4), major mode of transportation (1.1), and frequency of GSV use (1.9). Small but statistically significant differences were observed among age groups (4.6), time spent completing the task

Table 4
Average percent of agreement between crowdworkers and trained auditors.

	Street (1)	Street (2)	Street (3)	All streets (average)
Sidewalks	100.0	85.7	85.7	90.4
Wide sidewalks	98.0	93.3	93.0	94.8
Obstructions	78.0	59.0	17.7	51.6
Steep slopes	100.0	27.0	99.7	75.6
Street parking	92.7	78.3	97.7	89.6
Heavy traffic	64.0	99.7	96.7	86.8
Heavy foot traffic	97.7	99.7	99.3	98.9
Crosswalks	87.0	97.0	71.3	85.1
Traffic mirrors	95.0	99.0	96.7	96.9
Streetlights	92.3	53.0	49.3	64.9
Street trees	87.7	86.3	93.7	89.2
Attractive streetscape	30.7	77.7	94.3	67.6
Graffiti and litter	52.3	98.7	99.0	83.3
Abandoned buildings	97.0	98.7	97.3	97.7
All items (average)	83.7	82.4	85.1	83.7

Table 5
Inter-rater agreement of sub-total and total scores between crowdworkers and trained auditors.

	ICC[a]
Physical condition score	0.69
Safety score	0.58
Aesthetic score	0.30
Total score	0.65

[a] A two-way mixed model ICC, absolute agreement, single measures.

Table 6
Differences in the average percentage of agreement by crowdworker characteristics.

	n	%	# agreement	% agreement[a]
Total	300	100.0	35.2	83.7
Age				**
20s	57	19.1	34.0	81.0
30s	117	39.3	35.5	84.4
40s	97	32.6	35.2	83.9
50s	27	9.1	36.0	85.6
Gender				
Male	167	56.2	35.2	83.9
Female	130	43.8	35.1	83.5
Location of residence				
Urban center/downtown	87	29.0	34.7	82.6
Suburb/residential area	196	65.3	35.4	84.2
Rural area	17	5.7	35.3	84.0
History of residence in Nagoya				
Yes	280	93.3	35.1	83.6
No	20	6.7	35.7	85.0
Mode of transportation				
On foot	56	19.0	34.9	83.1
Bicycle	47	15.9	35.0	83.3
Train/bus	61	20.7	35.1	83.7
Car	131	44.4	35.3	84.1
Time spent on the task				**
< 15 min	111	37.0	34.6	82.3
15–29 min	148	49.3	35.4	84.3
≥ 30 min	41	13.7	36.0	85.6
Device used for the task				***
PC	252	84.0	35.5	84.4
Smartphone/tablet	48	16.0	33.6	80.0
Frequency of the use of GSV				
Often	68	22.7	34.6	82.5
Sometimes	170	56.7	35.3	84.0
Rarely/none	62	20.7	35.4	84.3

***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, +: $p < 0.1$.
[a] Differences were tested using an ANOVA or *t*-test.

(3.3), and type of electronic device used for the task (4.4). Those who were younger, spent less time completing the task, and used a smartphone/tablet for the task tended to have lower percentages of agreement.

4. Discussion

As the first study to investigate the efficiency and reliability of virtual neighborhood walkability audits conducted by a large number of untrained crowdworkers, the results of this study will support further work in this area. Crowdsourced virtual audits appear to help expand the target area, addressing one of the method's biggest limitations by reducing temporal and economic costs and enabling large-scale recruitment. In addition, although the use of virtual audit has been largely limited to the United States and other Western countries (Rzotkiewicz et al., 2018), this study showed that the GSV audit can be applicable to Japan, which has different urban forms and different availability of GSV imagery, indicating the generalizability regarding the usefulness of this method. As GSV or other similar online services have become increasingly available in many cities including those in non-Western countries, and crowdsourcing platforms are available in/accessible from many countries, crowdsourced virtual audits have a large potential to contribute to future research.

Using visual instructions and a brief and context-specific checklist for the virtual audit was shown to be efficient and reliable. In line with many previous studies' findings (Rzotkiewicz et al., 2018), GSV use reduced the time spent auditing streets compared to in-person audits as well as eliminated travel time to and between target streets. This study reaffirms the notion that virtual audits can substantially reduce the temporal costs of observation.

Also, agreement between the two trained auditors and between in-person and virtual audit methods was generally high. ICCs of total scores were both evaluated as excellent (ICC≧0.75). However, inter-rater reliability was low for the aesthetic score (ICC = 0.40) and for some of its components (e.g., graffiti and litter, Kappa = 0.06). These findings are also in line with previous studies that indicate that subjective items such as attractive streetscapes or small features such as discarded cigarettes are difficult to rate accurately (Clarke et al., 2010; Gullón et al., 2015; Rzotkiewicz et al., 2018). Inter-source reliability was relatively low for parking, traffic, and graffiti and litter, indicating that features subject to change over short periods of time tend to show lower agreement across audit methods (Clarke et al., 2010) due to time gaps between the dates of in-person audit and GSV imagery captured (an average of one to two years in this study). Additionally, detailed features on sidewalk may be challenged by the shooting angle of GSV (Aghaabbasi et al., 2018), which could have contributed to the low inter-source reliability of graffiti and litter. Although overall reliability was found to be sufficiently high, further research should continue refining checklist items and audit instructions.

Although simple checklists and instructions may be limited in their ability to accurately measure diverse streetscapes, they can be used for both practical and research purposes (Brownson et al., 2009; Cain et al., 2017). In practice, these tools may serve as a gateway to increase awareness among researchers, policymakers, and citizens about the importance and ways of measuring walkability. Additionally, some elements of the micro-scale streetscape may be easier to modify (e.g., installing street lights and removing graffiti) compared to elements of macro-scale walkability, such as increasing population density. This distinction is especially important for public health and urban planning policymakers to acknowledge when considering ways to increase the walkability of their city.

For research purposes, simple audit tools appear crucial to expanding survey areas, as they reduce the time spent auditing streets. Traditional in-person audits can only be conducted by a limited number of trained auditors, making this approach more time and resource intensive. However, simple audit tools combined with street imagery tools such as GSV and crowdsourcing allows many individuals to participate in auditing neighborhoods remotely. Since virtual audits by crowdworkers may be limited in their ability to provide intensive auditor training, concerns regarding the reliability among diverse crowdworkers persist. However, our analysis of 300 crowdworkers' audits found that their ratings were largely reliable, with only small differences arising as a function of their individual attributes.

In total, the percentage of agreement between the untrained crowdworkers' ratings and the trained auditors' ratings was very good (≧ 80%), and ICC for the total score was shown to have good reliability (≧ 0.60). This suggests that most of the crowdworkers rated street walkability in a way similar to that of trained auditors, relying only on the brief visual instructions in the absence of intensive on-site training. However, those reliability measures for crowdworkers were shown to be slightly lower when compared to the agreement between the two trained auditors' ratings for 830 street segments. Close examination of each item showed that some items had very low percentage of agreement values. For example, the majority of the crowdworkers gave different ratings from the trained auditors regarding steep slopes in street

2 and for obstructions in street 3. In these cases, it seemed to be difficult for those taking a casual look at the street to assess these items using street view imagery. The lower percentage of agreement for obstructions and attractive streetscape may be driven by the subjective nature of evaluating such items, an issue of perception that also occurred with the trained auditors. Although the low agreement for streetlights in streets 2 and 3 were unexpected, these items proved to be erratic cases in which streetlights were located immediately above the starting point, were very small, or were installed on electric poles, making them less visible. This low agreement could be partially improved if more detailed instructions or conducting online training was provided, although such additional training may increase study costs.

Another important finding regarding crowdworker audits was that their ratings were largely unrelated to individual attributes including gender, residential location, history of residence, and mode of transportation. As with the results of a previous study by Zhu et al. (2017), which reported fair-to-substantial reliability across raters with different familiarity with a region, neither location of current residence nor history of residence were associated with the accuracy of ratings. The frequency of GSV use was also found to be unrelated to the ratings, suggesting that the inexperienced crowdworkers did not experience operational difficulties in auditing using GSV. These results support the use of GSV audits by crowdworkers, as the method enables many untrained auditors to participate in observation remotely, regardless of their residential characteristics and experience with GSV.

Age proved to be the only individual attribute associated with the reliability of the ratings. The percentage of agreement was lower for those in their 20s than other age groups, although the difference was less than 5% and was found to be related to the device used for the task. Younger crowdworkers tended to use smartphones for the audit, and those who used smartphones produced ratings with lower percentage of agreement values. This finding is notable because the use of smartphones for crowdworking may be increasing. However, the effect of smartphones can be controlled by specifying the type of device to be used for audits. In addition, since shorter times to complete the task was also related to lower percentage of agreement, paid-per-time crowdsourcing with upper limits may improve reliability by decreasing incentives for quick completion. Thus, while this study found some factors negatively associated with inter-rater reliability among crowdworkers, this negative effect can be managed to a large extent.

Importantly, this study has several limitations. First, the number of street segments for the crowdworker's audit was limited to three. This was because our project prioritized recruiting more crowdworkers over street segments, and a small task which can be completed in a short time was considered suitable for recruiting many crowdworkers. However, the small number of street segments limited our analysis substantially: for example, the Kappa coefficient for each item was not computed because it could yield very unstable values and was considered inappropriate. Therefore, the results of this study are preliminary and further research targeting a large number of street segments is required to expand knowledge about the use of crowdsourcing for virtual audits. Second, further studies are needed to refine the checklist. For example, we regarded the absence of traffic mirrors as indicative of safe environments, as there appeared no need to install such mirrors, but this might actually reflect a deficit of investment in neighborhood safety. Numerical values or objective criteria (e.g., the number of pedestrians for “heavy foot traffic”) could be considered for some items as long as simplicity is maintained for the auditors. Third, more reliability tests for the tool are necessary. Although we only tested inter-rater reliability for the virtual audit, the test should be extended to in-person audit, using data of two or more raters auditing on the same day. Investigating the order effect of in-person and virtual audit remain overlooked in the current study. Fourth, this study used *Lancers* for recruiting crowdworkers. Future research should compare *Lancers* to other crowdsourcing platforms such as Amazon Mechanical Turk or unpaid, voluntary crowdsourcing to confirm whether any biases exist among participants from different sources. Finally, while only the latest imagery from GSV was used in this study, GSV has the potential to track the neighborhood changes over time. However, there are challenges when using GSV for such research purposes: availability of GSV imagery and its frequency of updates are not uniform across cities, neighborhoods, or even street segments, and information on the day of the week and time of the day are not currently available, which may cause difficulty in comparing some items equally (e.g., traffic volume). Virtually identifying and measuring changes in neighborhood streetscape and analyzing their influence on health behaviors should be explored in future studies.

5. Conclusion

This study examined the efficiency and reliability of virtual audits of neighborhood walkability, using visual instructions and a simple checklist developed for the purposes of this study. Analysis for the trained auditors' ratings established virtual audits as efficient and reliable. Additionally, untrained crowdworkers’ ratings exhibited good agreement with those of trained auditors, and differences in ratings observed across individual attributes, experience of GSV, and work environment were largely small and manageable, if any. In conclusion, auditing neighborhoods virtually through the use of crowdsourcing has large potential to expand study areas while keeping various audit items, therefore addressing the methodological limits of audits by trained auditors and computer vision approaches that are emerging in neighborhood and health studies.

Declarations of interest

None.

Acknowledgements

This work was supported by DAIKO FOUNDATION, Japan (Grant Number 11047) and JSPS KAKENHI, Japan (Grant Numbers JP17H00947, JP15H02964, and JP18KK0371).

References

Aghaabbasi, M., Moeinaddini, M., Shah, M.Z., Asadi-Shekari, Z., 2018. Addressing issues in the use of Google tools for assessing pedestrian built environments. J. Transp. Geogr. 73, 185–198.

Badland, H.M., Opit, S., Witten, K., Kearns, R.A., Mavoa, S., 2010. Can virtual streetscape audits reliably replace physical streetscape audits? J. Urban Health 87 (6), 1007–1016.

Ben-Joseph, E., Lee, J.S., Cromley, E.K., Laden, F., Troped, P.J., 2013. Virtual and actual: relative accuracy of on-site and web-based instruments in auditing the environment for physical activity. Health Place 19, 138–150.

Brownson, R.C., Hoehner, C.M., Day, K., Forsyth, A., Sallis, J.F., 2009. Measuring the built environment for physical activity: state of the science. Am. J. Prev. Med. 36 (4 Suppl. l), S99–S123 e12.

Cain, K.L., Gavand, K.A., Conway, T.L., Geremia, C.M., Millstein, R.A., Frank, L.D., Saelens, B.E., Adams, M.A., Glanz, K., King, A.C., Sallis, J.F., 2017. Developing and validating an abbreviated version of the microscale audit for pedestrian streetscapes (MAPS-Abbreviated). J. Transp. Health 5, 84–96.

Cerin, E., Chan, K.W., Macfarlane, D.J., Lee, K.Y., Lai, P.C., 2011. Objective assessment of walking environments in ultra-dense cities: development and reliability of the Environment in Asia Scan Tool–Hong Kong version (EAST-HK). Health Place 17 (4), 937–945.

Cicchetti, D.V., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess. 6 (4), 284–290.

Clarke, P., Ailshire, J., Melendez, R., Bader, M., Morenoff, J., 2010. Using Google Earth to conduct a neighborhood audit: reliability of a virtual audit instrument. Health Place 16 (6), 1224–1229.

Clifton, K.J., Livi Smith, A.D., Rodriguez, D., 2007. The development and testing of an audit for the pedestrian environment. Landsc. Urban Plan. 80 (1–2), 95–110.

Diez Roux, A.V., 2007. Neighborhoods and health: where are we and were do we go from here? Revue d'Épidémiologie et de Santé Publique 55 (1), 13–21.

Ding, D., Gebel, K., 2012. Built environment, physical activity, and obesity: what have we learned from reviewing the literature? Health Place 18 (1), 100–105.

Ferdinand, A.O., Sen, B., Rahurkar, S., Engler, S., Menachemi, N., 2012. The relationship between built environments and physical activity: a systematic review. Am. J. Public

Health 102 (10), e7–e13.
Grasser, G., Van Dyck, D., Titze, S., Stronegger, W., 2013. Objectively measured walkability and active transport and weight-related outcomes in adults: a systematic review. Int. J. Public Health 58 (4), 615–625.
Gullón, P., Badland, H.M., Alfayate, S., Bilal, U., Escobar, F., Cebrecos, A., Diez, J., Franco, M., 2015. Assessing walking and cycling environments in the streets of Madrid: comparing on-field and virtual audits. J. Urban Health 92 (5), 923–939.
Hara, K., Le, V., Froehlich, J., 2013. Combining crowdsourcing and google street view to identify street-level accessibility problems. In: CHI '13 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 631–640.
Hipp, J.A., Adlakha, D., Eyler, A.A., Gernes, R., Kargol, A., Stylianou, A.H., Pless, R., 2017. Learning from outdoor webcams: surveillance of physical activity across environments. In: Thakuriah, P., Tilahun, N., Zellner, M. (Eds.), Seeing Cities through Big Data. Springer, New York, pp. 471–490.
Hoedl, S., Titze, S., Oja, P., 2010. The bikeability and walkability evaluation table reliability and application. Am. J. Prev. Med. 39 (5), 457–459.
Kaido, K., 2006. Urban densities, quality of life and local facility accessibility in principal Japanese cities. In: Jenks, M., Dempsey, N. (Eds.), Future Forms and Design for Sustainable Cities. Architectural Press, Oxford, pp. 311–337.
Kavanagh, C., Thomson, R., Yuki, M., 2016. A critical survey of online users of a popular Japanese crowdsourcing website. In: [Poster]. 23rd Congress of the International Association for Cross Cultural Psychology, Nagoya, Japan.
Kelly, C.M., Wilson, J.S., Baker, E.A., Miller, D.K., Schootman, M., 2013. Using Google Street View to audit the built environment: inter-rater reliability results. Ann. Behav. Med. 45 (Suppl. 1), S108–S112.
Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174.
Malecki, K.C., Engelman, C.D., Peppard, P.E., Nieto, F.J., Grabow, M.L., Bernardinello, M., Bailey, E., Bersch, A.J., Walsh, M.C., Lo, J.Y., Martinez-Donate, A., 2014. The Wisconsin Assessment of the Social and Built Environment (WASABE): a multi-dimensional objective audit instrument for examining neighborhood effects on health. BMC Public Health 14, 1165.
Nguyen, Q.C., Sajjadi, M., McCullough, M., Pham, M., Nguyen, T.T., Yu, W., Meng, H.W., Wen, M., Li, F., Smith, K.R., Brunisholz, K., Tasdizen, T., 2018. Neighbourhood looking glass: 360° automated characterisation of the built environment for neighbourhood effects research. J. Epidemiol. Community Health 72 (3), 260–266.
Pliakas, T., Hawkesworth, S., Silverwood, R.J., Nanchahal, K., Grundy, C., Armstrong, B., Casas, J.P., Morris, R.W., Wilkinson, P., Lock, K., 2017. Optimising measurement of health-related characteristics of the built environment: comparing data collected by foot-based street audits, virtual street audits and routine secondary data sources. Health Place 43, 75–84.
Rundle, A.G., Bader, M.D., Richards, C.A., Neckerman, K.M., Teitler, J.O., 2011. Using google street view to audit neighborhood environments. Am. J. Prev. Med. 40 (1), 94–100.
Rzotkiewicz, A., Pearson, A.L., Dougherty, B.V., Shortridge, A., Wilson, N., 2018. Systematic review of the use of Google Street View in health research: major themes, strengths, weaknesses and possibilities for future research. Health Place 52, 240–246.
Saelens, B.E., Handy, S.L., 2008. Built environment correlates of walking: a review. Med. Sci. Sport. Exerc. 40 (7 Suppl. l), S550–S566.
Sallis, J.F., Cain, K.L., Conway, T.L., Gavand, K.A., Millstein, R.A., Geremia, C.M., Frank, L.D., Saelens, B.E., Glanz, K., King, A.C., 2015. Is your neighborhood designed to support physical activity? A brief streetscape audit tool. Prev. Chronic Dis. 12, E141.
Sallis, J.F., Cerin, E., Conway, T.L., Adams, M.A., Frank, L.D., Pratt, M., Salvo, D., Schipperijn, J., Smith, G., Cain, K.L., Davey, R., Kerr, J., Lai, P.C., Mitáš, J., Reis, R., Sarmiento, O.L., Schofield, G., Troelsen, J., Van Dyck, D., De Bourdeaudhuij, I., Owen, N., 2016. Physical activity in relation to urban environments in 14 cities worldwide: a cross-sectional study. Lancet 387 (10034), 2207–2217.
Schaefer-McDaniel, N., Caughy, M.O., O'Campo, P., Gearey, W., 2010. Examining methodological details of neighbourhood observations and the relationship to health: a literature review. Soc. Sci. Med. 70 (2), 277–292.
Shelton, B., 2012. Learning from the Japanese City: Looking East in Urban Design. Routledge, London.
Su, M., Du, Y.K., Liu, Q.M., Ren, Y.J., Kawachi, I., Lv, J., Li, L.M., 2014. Objective assessment of urban built environment related to physical activity–development, reliability and validity of the China Urban Built Environment Scan Tool (CUBEST). BMC Public Health 14, 109.
Zhu, W., Sun, Y., Kurka, J., Geremia, C., Engelberg, J.K., Cain, K., Conway, T., Sallis, J.F., Hooker, S.P., Adams, M.A., 2017. Reliability between online raters with varying familiarities of a region: microscale Audit of Pedestrian Streetscapes (MAPS). Landsc. Urban Plan. 167, 240–248.